MIT 9.520/6.7910,
Statistical Learning Theory and Applications

Class 03: Least squares and overpamaterization

Lorenzo Rosasco

Learning problem and algorithms

Solve

$$\min_{f \in \mathcal{F}} L(f) = L(f_P), \qquad L(f) = \mathbb{E}_{(x,y)\sim P}[\ell(y, f(x))],$$

given only

$$S_n = (x_1, y_1), \ldots, (x_n, y_n) \sim P^n.$$

Learning problem and algorithms

Solve

$$\min_{f \in \mathcal{F}} L(f) = L(f_P), \qquad L(f) = \mathbb{E}_{(x,y)\sim P}[\ell(y, f(x))],$$

given only

$$S_n = (x_1, y_1), \ldots, (x_n, y_n) \sim P^n.$$

Learning algorithm

$$S_n \to \widehat{f} = \widehat{f}_{S_n},$$

$\widehat{f}$ estimates f_P given the observed examples S_n.

Learning problem and algorithms

Solve

$$\min_{f \in \mathcal{F}} L(f) = L(f_P), \qquad L(f) = \mathbb{E}_{(x,y)\sim P}[\ell(y, f(x))],$$

given only

$$S_n = (x_1, y_1), \ldots, (x_n, y_n) \sim P^n.$$

Learning algorithm

$$S_n \to \widehat{f} = \widehat{f}_{S_n},$$

$\widehat{f}$ estimates f_P given the observed examples S_n.

How can we design a learning algorithm?

Classic algorithm design: representation

Pick a model, e.g. parameterized

Linearly

$$w^T x \quad \mapsto \quad w^\top \Phi(x)$$

Classic algorithm design: representation

Pick a model, e.g. parameterized

Linearly

$$w^T x \quad \mapsto \quad w^\top \Phi(x)$$

or

Non linearly

$$w^T \sigma(Wx) \quad \mapsto \quad w^T \sigma(W_L \dots \sigma W_2 \sigma(W_1 x)$$

Classic algorithm design: representation

Pick a model, e.g. parameterized

Linearly

$$w^T x \quad \mapsto \quad w^\top \Phi(x)$$

or

Non linearly

$$w^T \sigma(Wx) \quad \mapsto \quad w^T \sigma(W_L \dots \sigma W_2 \sigma(W_1 x)$$

This includes various choices like features/kernels, architecture and possible constraints e.g.

$$\|w\| \leq R, \quad \|W_j\| \leq R_j.$$

Classic algorithm design: optimization

Given some model, e.g. $f_w(x), \quad \|w\| \leq R$, consider the empirical risk minimization (ERM)

$$\min_{\|w\| \leq R} \widehat{L}(f_w) \qquad \widehat{L}(f_w) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_w(x_i)),$$

or

$$\min_{w} \widehat{L}(f_w) + \lambda \|w\| .$$

Outline

Overparameterization and bias

Regularization and stability

Representer theorem and numerical complexity

Beyond ℓ^2 regularization

Linear functions and linear spaces

Let $\mathcal{H}$ be the space of linear functions

$$f(x) = w^\top x.$$

Then,

- $f \leftrightarrow w$ is one to one,
- inner product $\langle f, \bar{f} \rangle_{\mathcal{H}} := w^\top \bar{w}$,
- norm/metric $\left\| f - \bar{f} \right\|_{\mathcal{H}} := \|w - \bar{w}\|$.

Linear functions are the conceptual building block of most functions class.

Linear least squares

ERM with least squares also called ordinary least squares (OLS)

$$\min_{w \in \mathbb{R}^d} \underbrace{\frac{1}{n} \sum_{i=1}^{n} (y_i - w^\top x_i)^2}_{\widehat{L}(w)}.$$

Discussing computations provides many insights.

Matrices and linear systems

Let $\widehat{X} \in \mathbb{R}^{nd}$ and $\widehat{Y} \in \mathbb{R}^n$. Then

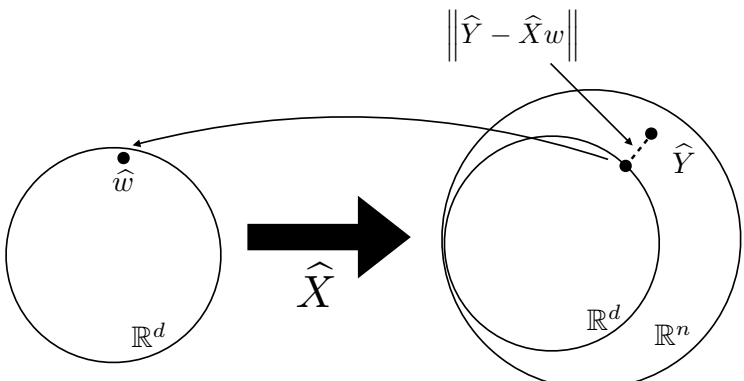
$$\frac{1}{n}\sum_{i=1}^{n}(y_i - w^\top x_i)^2 = \frac{1}{n}\left\|\widehat{Y} - \widehat{X}w\right\|^2.$$

This is the least squares problem associated to the linear system

$$\widehat{X}w = \widehat{Y}.$$

Overdetermined lin. syst.
aka underparameterized models

$n > d$



Least squares solutions

From the optimality conditions

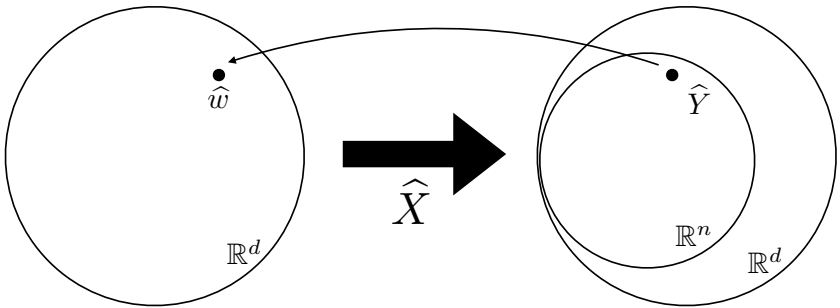
$$\nabla_w \frac{1}{n} \left\| \widehat{Y} - \widehat{X} w \right\|^2 = 0$$

we can derive the normal equation

$$\widehat{X}^\top \widehat{X} w = \widehat{X}^\top \widehat{Y} \qquad \Leftrightarrow \qquad \widehat{w} = (\widehat{X}^\top \widehat{X})^{-1} \widehat{X}^\top \widehat{Y}.$$

Underdetermined lin. syst.
aka overparameterized model

$n < d$



Minimal norm solution

There can be many solutions

$$\widehat{X}\widehat{w} = \widehat{Y}, \quad \text{and} \quad \widehat{X}w_0 = 0 \quad \Rightarrow \widehat{X}(\widehat{w} + w_0) = \widehat{Y}.$$

Consider

$$\min_{w \in \mathbb{R}^d} \|w\|^2, \qquad \text{subj. to} \qquad \widehat{X}w = \widehat{Y}.$$

Using the method of Lagrange multipliers, the solution is

$$\widehat{w} = \widehat{X}^\top (\widehat{X}\widehat{X}^\top)^{-1}\widehat{Y}.$$

Computing the minimal norm solution

$$\min_{w \in \mathbb{R}^d} \|w\|^2 \quad \text{subj.t.} \quad \widehat{X} w = \widehat{Y}.$$

- Lagrangian
$$\mathcal{L}(w, \lambda) = \|w\|^2 + \lambda^\top (\widehat{Y} - \widehat{X} w).$$
- First-order condition
$$\nabla_w \mathcal{L} = 2w - \widehat{X}^\top \lambda = 0 \quad \Rightarrow \quad w = \tfrac{1}{2} \widehat{X}^\top \lambda.$$
- Substitute into the constraint
$$\widehat{X} w = \tfrac{1}{2} \widehat{X} \widehat{X}^\top \lambda = \widehat{Y} \quad \Rightarrow \quad (\widehat{X} \widehat{X}^\top) \lambda = 2 \widehat{Y}.$$
- Finally
$$\lambda = 2 (\widehat{X} \widehat{X}^\top)^{-1} \widehat{Y}, \quad \widehat{w} = \widehat{X}^\top (\widehat{X} \widehat{X}^\top)^{-1} \widehat{Y}.$$

Geometric view on minimal norm

$$\widehat{w} = \widehat{X}^\top (\widehat{X}\widehat{X}^\top)^{-1}\widehat{Y} \quad \Rightarrow \quad \widehat{w} \in \mathrm{Ran}(\widehat{X}^\top)$$

Recall the orthogonal decomposition

$$\mathbb{R}^d = \mathrm{Ran}(\widehat{X}^\top) \oplus \mathrm{Ker}(\widehat{X}).$$

Any solution of $\widehat{X}w = \widehat{Y}$ can be written as

$$w = \widehat{w} + w_0, \quad w_0 \in \mathrm{Ker}(\widehat{X}),$$

that is the minimal norm solution $\widehat{w}$ is the unique one orthogonal to $\ker(\widehat{X})$,

Pseudoinverse

$$\widehat{w} = \widehat{X}^{\dagger}\widehat{Y}$$

For $n > d$, (independent columns)

$$\widehat{X}^{\dagger} = (\widehat{X}^{\top}\widehat{X})^{-1}\widehat{X}^{\top}.$$

For $n < d$, (independent rows)

$$\widehat{X}^{\dagger} = \widehat{X}^{\top}(\widehat{X}\widehat{X}^{\top})^{-1}.$$

Spectral view

Consider the SVD of $\widehat{X}$

$$\widehat{X} = USV^\top \quad \Leftrightarrow \quad \widehat{X}w = \sum_{j=1}^{r} s_j (v_j^\top w) u_j,$$

here $r \leq n \wedge d$ is the rank of $\widehat{X}$.

Then,

$$\widehat{w}^\dagger = \widehat{X}^\dagger \widehat{Y} = \sum_{j=1}^{r} \frac{1}{s_j} (u_j^\top \widehat{Y}) v_j.$$

Pseudoinverse and bias

$$\widehat{w}^{\dagger} = \widehat{X}^{\dagger}\widehat{Y} = \sum_{j=1}^{r} \frac{1}{s_j}(u_j^{\top}\widehat{Y})v_j.$$

$(v_j)_j$ are principal components of $\widehat{X}$: OLS "likes" principal components.

Not all linear functions are the same for OLS!

The pseudoinverse introduces a bias towards certain solutions.

Outline

Overparameterization and bias

Regularization and stability

Representer theorem and numerical complexity

Beyond ℓ^2 regularization

From OLS to ridge regression

Recall, it also holds,

$$\widehat{X}^\dagger = \lim_{\lambda \to 0_+} (\widehat{X}^\top \widehat{X} + \lambda I)^{-1} \widehat{X}^\top = \lim_{\lambda \to 0_+} \widehat{X}^\top (\widehat{X}\widehat{X}^\top + \lambda I)^{-1}.$$

Consider for $\lambda > 0$,

$$\widehat{w}_\lambda = (\widehat{X}^\top \widehat{X} + \lambda I)^{-1} \widehat{X}^\top \widehat{Y}.$$

This is called ridge regression.

Spectral view on ridge regression

$$\widehat{w}_\lambda = (\widehat{X}^\top \widehat{X} + \lambda I)^{-1} \widehat{X}^\top \widehat{Y}$$

Considering the SVD of $\widehat{X}$,

$$\widehat{w}_\lambda = \sum_{j=1}^{r} \frac{s_j}{s_j^2 + \lambda} (u_j^\top \widehat{Y}) v_j.$$

Ridge regression as filtering

$$\widehat{w}_\lambda = \sum_{j=1}^{r} \frac{s_j}{s_j^2 + \lambda} (u_j^\top \widehat{Y}) v_j$$

The function

$$F(s) = \frac{s}{s^2 + \lambda},$$

acts as a low pass filter (low frequencies= principal components).

- For s small, $F(s) \approx 1/\lambda$.
- For s big, $F(s) \approx 1/s$.

Ridge regression as ERM

$$\widehat{w}_\lambda = (\widehat{X}^\top \widehat{X} + \lambda I)^{-1} \widehat{X}^\top \widehat{Y}$$

is the solution of

$$\min_{w \in \mathbb{R}^d} \underbrace{\left\| \widehat{Y} - \widehat{X} w \right\|^2 + \lambda \|w\|^2}_{\widehat{L}_\lambda(w)}.$$

It follows from,

$$\Delta \widehat{L}_\lambda(w) = -\frac{2}{n} \widehat{X}^\top (\widehat{Y} - \widehat{X} w) + 2\lambda w = 2(\frac{1}{n} \widehat{X}^\top \widehat{X} + \lambda I) w - \frac{2}{n} \widehat{X}^\top \widehat{Y}.$$

Ridge regression as ERM

ERM interpretation suggests the rescaling

$$\widehat{w}_\lambda = (\widehat{X}^\top \widehat{X} + n\lambda I)^{-1} \widehat{X}^\top \widehat{Y}$$

since

$$\min_{w \in \mathbb{R}^d} \underbrace{\frac{1}{n} \left\| \widehat{Y} - \widehat{X} w \right\|^2 + \lambda \|w\|^2}_{\widehat{L}_\lambda(w)}.$$

Related ideas

Tikhonov

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \left\| \widehat{Y} - \widehat{X} w \right\|^2 + \lambda \|w\|^2$$

Morozov

$$\min_{w \in \mathbb{R}^d} \|w\|^2 \qquad \text{subj. to} \qquad \frac{1}{n} \left\| \widehat{Y} - \widehat{X} w \right\|^2 \leq \delta$$

Ivanov

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \left\| \widehat{Y} - \widehat{X} w \right\|^2, \qquad \text{subj. to} \qquad \|w\|^2 \leq R$$

Note: equivalent optimization problems for appropriate parameter choices.

Ridge regression and structural risk minimization

The constraint

$$\|w\|^2 \leq R$$

- restricts the search of solution,
- shrinks the solution coefficients.

Different views on regularization

$$\widehat{w} = \widehat{X}^{\dagger}\widehat{Y} \qquad\qquad \widehat{w}_{\lambda} = (\widehat{X}^{\top}\widehat{X} + \lambda I)^{-1}\widehat{X}^{\top}\widehat{Y}$$

$$\min_{w \in \mathbb{R}^d \text{ s.t. } \widehat{X}w = \widehat{Y}} \|w\|^2 \qquad\qquad \min_{w \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}(y_i - w^{\top}x_i)^2 + \lambda \|w\|^2$$

- Introduces a *bias* towards certain solutions: small norm/principal components,
- controls the *stability* of the solution.

Outline

Overparameterization and bias

Regularization and stability

Representer theorem and numerical complexity

Beyond ℓ^2 regularization

Complexity of ridge regression

Back to computations.

Solving

$$\widehat{w}^{\lambda} = (\widehat{X}^{\top}\widehat{X} + \lambda I)^{-1}\widehat{X}^{\top}\widehat{Y}$$

requires essentially (using a direct solver)

- time $O(nd^2 + d^3)$,
- memory $O(nd \vee d^2)$.

What if $n \ll d$?

Representer theorem

A simple observation

We can show that (e.g. using SVD)

$$(\widehat{X}^\top \widehat{X} + \lambda I)^{-1} \widehat{X}^\top = \widehat{X}^\top (\widehat{X}\widehat{X}^\top + \lambda I)^{-1}$$

More on computational complexity

Then

$$\widehat{w}_\lambda = \widehat{X}^\top (\widehat{X}\widehat{X}^\top + \lambda I)^{-1} \widehat{Y}.$$

requires essentially (using a direct solver)

- time $O(n^2 d + n^3)$,
- memory $O(nd \vee n^2)$.

Representer theorem

Note that

$$\widehat{w}_\lambda = \widehat{X}^\top \underbrace{(\widehat{X}\widehat{X}^\top + \lambda I)^{-1}\widehat{Y}}_{c \in \mathbb{R}^n} = \sum_{i=1}^{n} x_i c_i.$$

The coefficients vector is a linear combination of the input points.

Then

$$\hat{f}_\lambda(x) = x^\top \widehat{w}_\lambda = x^\top \widehat{X}^\top c = \sum_{i=1}^{n} x^\top x_i c_i$$

The function we obtain is a linear combination of *inner products*.

This will be the key to nonparametric (kernel) learning.

Outline

Overparameterization and bias

Regularization and stability

Representer theorem and numerical complexity

Beyond ℓ^2 regularization

Other regularization terms

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - w^\top x_i)^2 + \lambda R(w)$$

Beyond ℓ^2, other regularization terms can be considered.

Popular choices are sparsity-inducing regularizers, e.g.

- ℓ^0 "norm":
$$R(w) = \|w\|_0 = \#\{j : w_j \neq 0\},$$
- ℓ^1 norm:
$$R(w) = \|w\|_1 = \sum_j |w_j|.$$

Why sparsity?

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \|\widehat{Y} - \widehat{X} w\|^2 + \lambda \|w\|_0$$

Advantages:
- Variable selection
- Interpretability

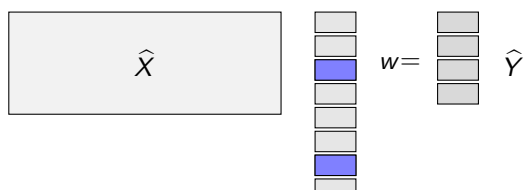
Challenges:
- Nonconvex optimization
- NP-hard in general

Sparsity and overparametrization

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \|\widehat{Y} - \widehat{X} w\|^2 + \lambda \|w\|_0$$

$$\min_{w \in \mathbb{R}^d} \|w\|_0$$

$$\text{s.t.} \quad \widehat{X} w = \widehat{Y}$$

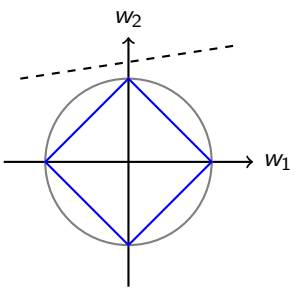


Why ℓ^1

$$\|w\|_0 \;\mapsto\; \|w\|_1 \neq \|w\|$$

$$\min_{w \in \mathbb{R}^d} \|w\|_1$$

$$\text{s.t.} \quad \widehat{X} w = \widehat{Y}$$



Computing the Lasso/Basis pursuit

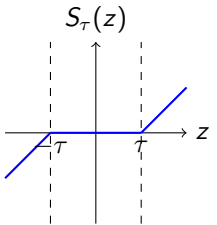
$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \|\widehat{Y} - \widehat{X} w\|^2 + \lambda \|w\|_1$$

ISTA: Iterative soft-thresholding

$$w^{(k+1)} = S_{\lambda\eta}\Big(w^{(k)} - \eta\, \widehat{X}^\top (\widehat{X} w^{(k)} - \widehat{Y})\Big)$$

Soft-thresholding

$$S_\tau(z) = \mathrm{sign}(z)\ \max\{|z| - \tau,\, 0\}$$



Remarks

- Other sparse regularizers: group lasso, TV-norm ...
- Other optimizers: Franke-Wolfe/conditional gradient (matching pursuit).
- No representer theorem.

Summing up

- ERM.
- From OLS to ridge regression.
- Different views: (spectral) filtering and ERM.
- Regularization and bias.
- Sparse regularization.